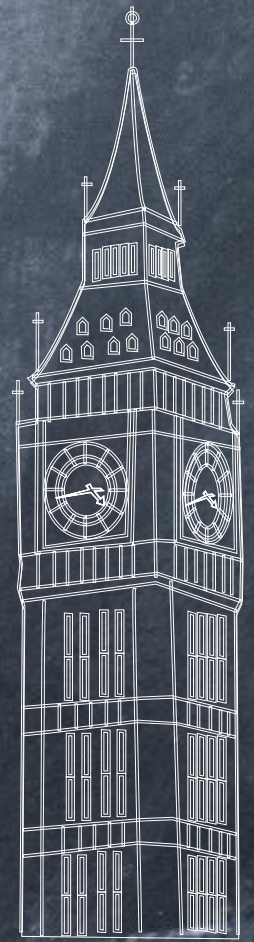
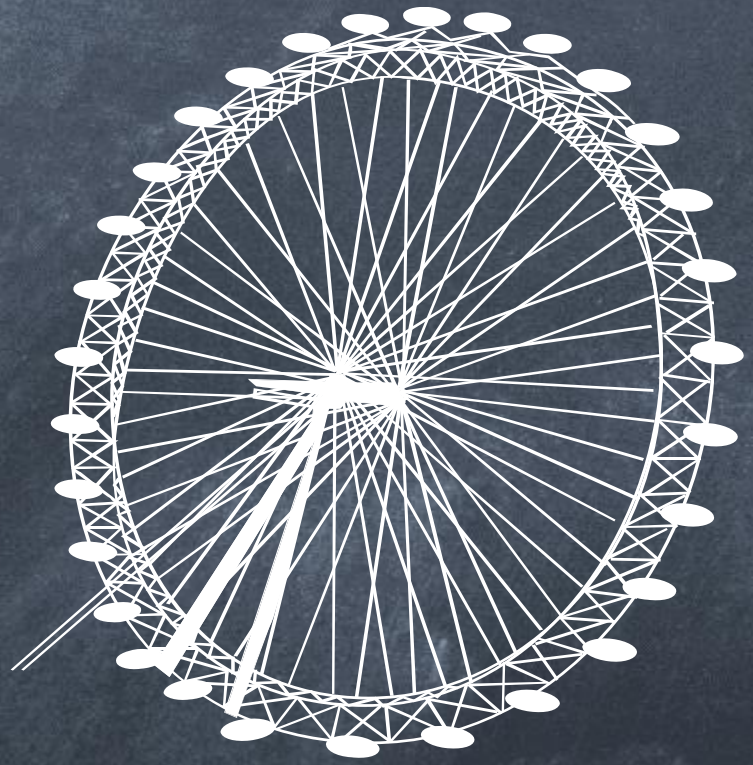
移动智能设备传感器安全研究

——通过手机传感器获取用户输入信息

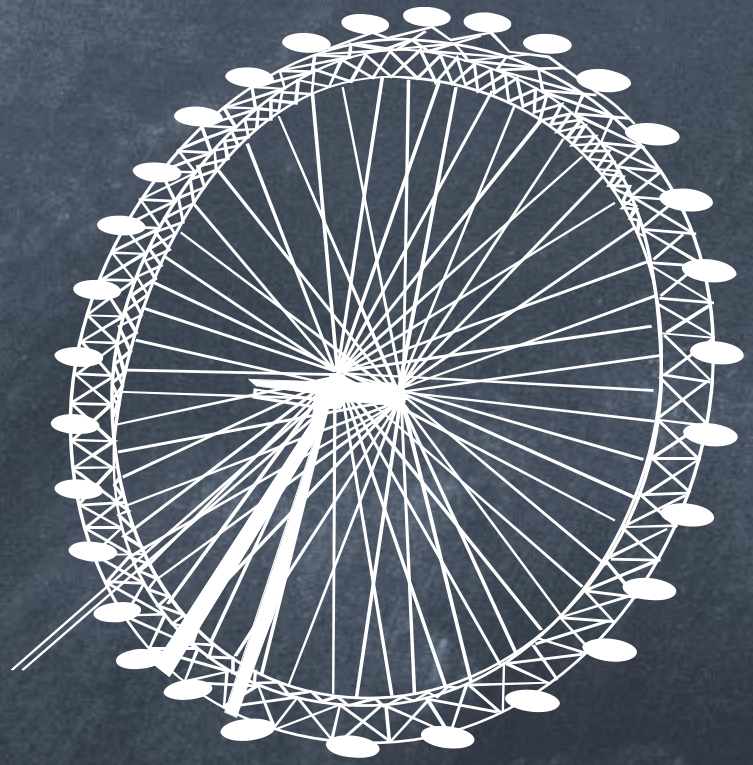
答辩人：朱孟垚





目录

Contents



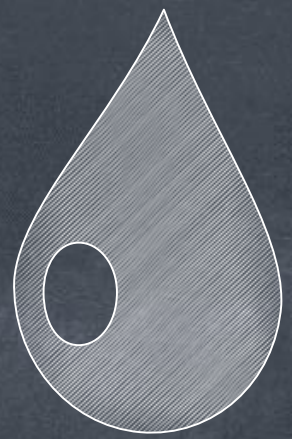
Part 01

选题的背景与意义

Background And Significance

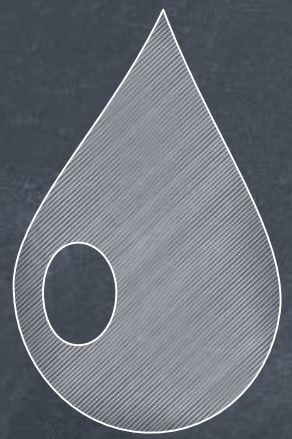
01.选题的背景与意义

Background And Significance



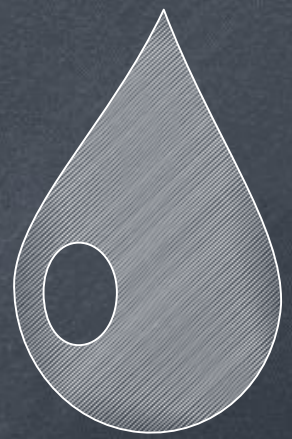
移动设备的普及

移动设备作为互联网生态圈中重要的组成部分，其承担着大多数使用者日常使用，身份验证，移动支付等众多需求。



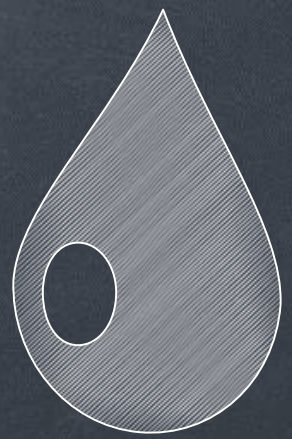
传感器的普及

加速度传感器和角度传感器在智能手机上十分普遍，上至旗舰机下到千元机。而且传感器的性能相当可靠



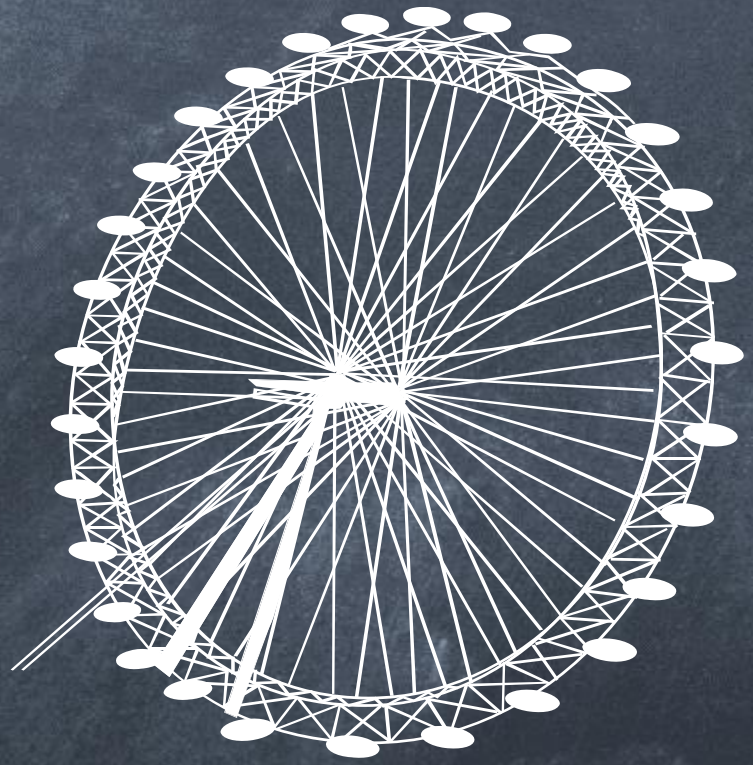
固有缺陷

用户在对智能机操作时，设备会有特定角度进行晃动，而人体对于角度的感知相比如其他感知能力是误差是较大的（头部为6°，躯干为3°），而手机的传感器足够采集到这些被忽视的晃动。



难以防治

目前对于旁信道攻击与防御相对不太被关注。而且这种攻击方式不利用常规意义上的漏洞，较为隐蔽。



Part 02

课题研究的思路

Ideas And Methods

02.课题研究的思路

Ideas And Methods

获取用户传感器信息，
并上传到主机

主机进行数据预处理，
找到有效区域数据

根据有效数据确定当前
输入模式

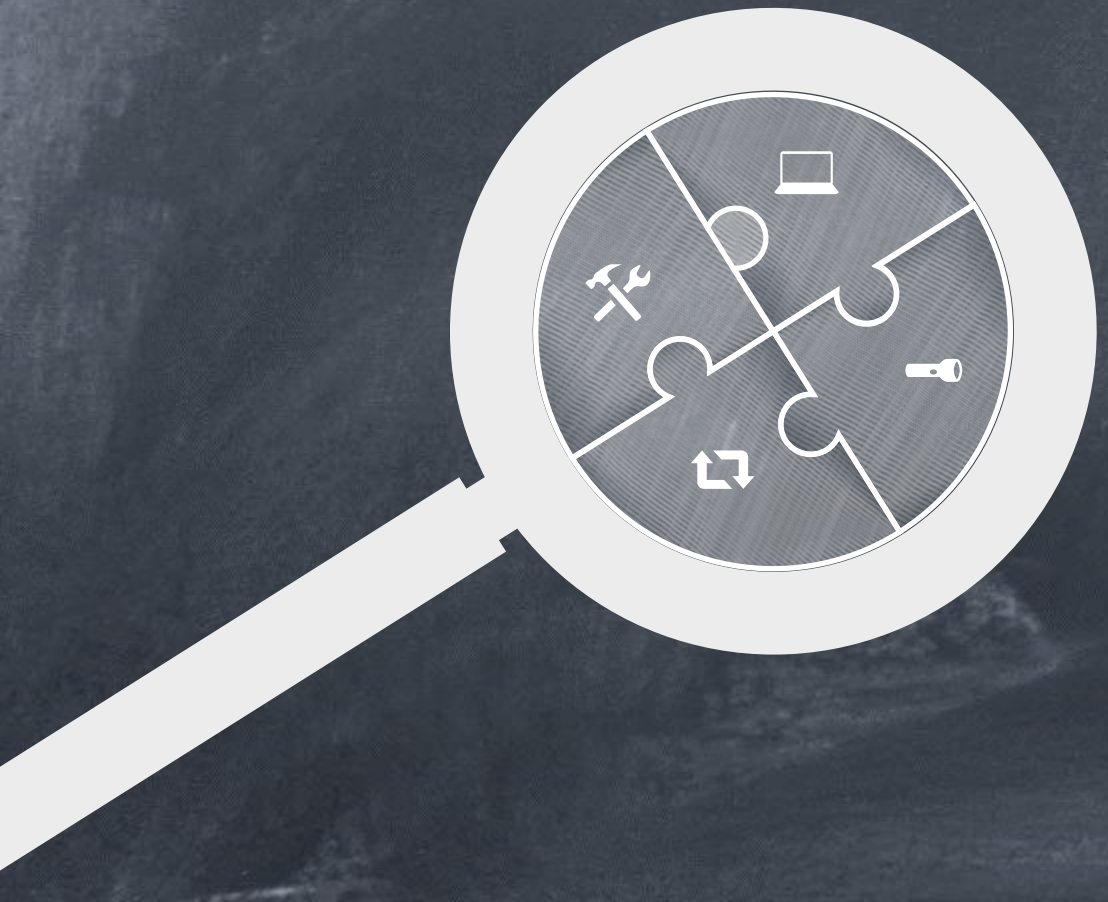
对有效区域进行分析，
将可能是文本输入的数
据进行预处理

通过模式匹配算法给出
预测

攻击过程思路

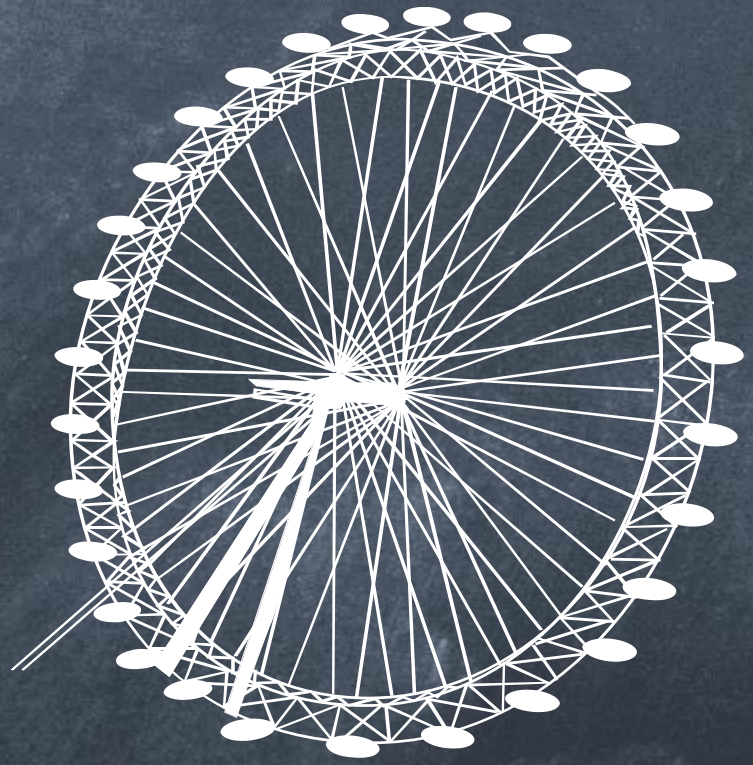
02.课题研究的思路

Ideas And Methods



技术实现研究思路

1，制作数据采集工具
2，收集数据并调试采集工具
3，对数据进行试预处理并且同时优化数据采集工具
4，确定数据标准采集工具与标准数据格式
5，归类有效数据，制作训练集
6，模拟日常使用情况，对比日常数据与标准数据的差距，
7，构建算法将日常使用中的数据提取出有效数据
8，构建可以进行模式匹配的算法 （ongoing）



Part 03

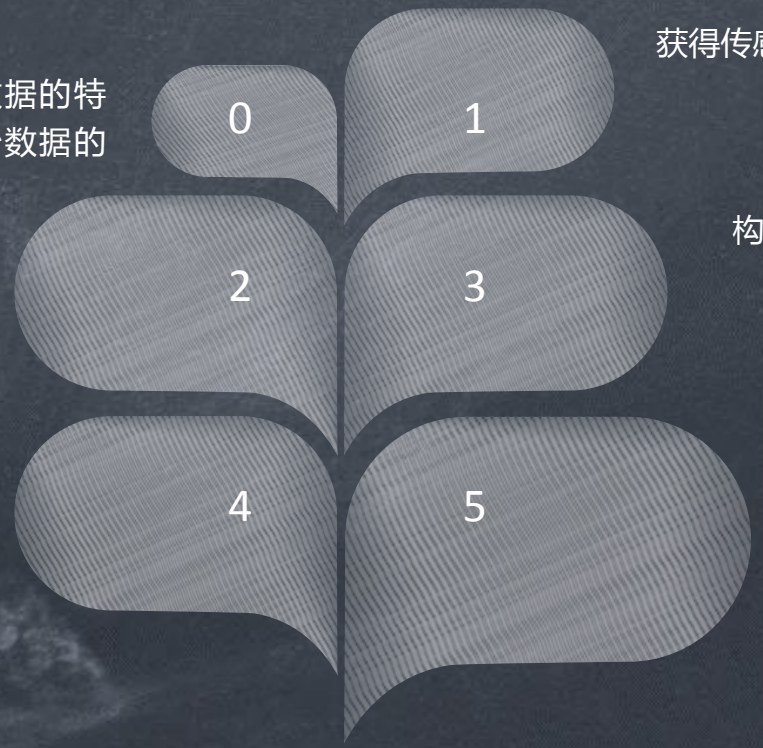
课题研究的进展

The Research Progress

03.课题研究的进展

The Research Progress

研究中需要的几个关键技术节点



传感器与原始数据的特点以及处理原始数据的算法

获得传感器数据并传递

预处理数据算法

构建不同模式下的数据库

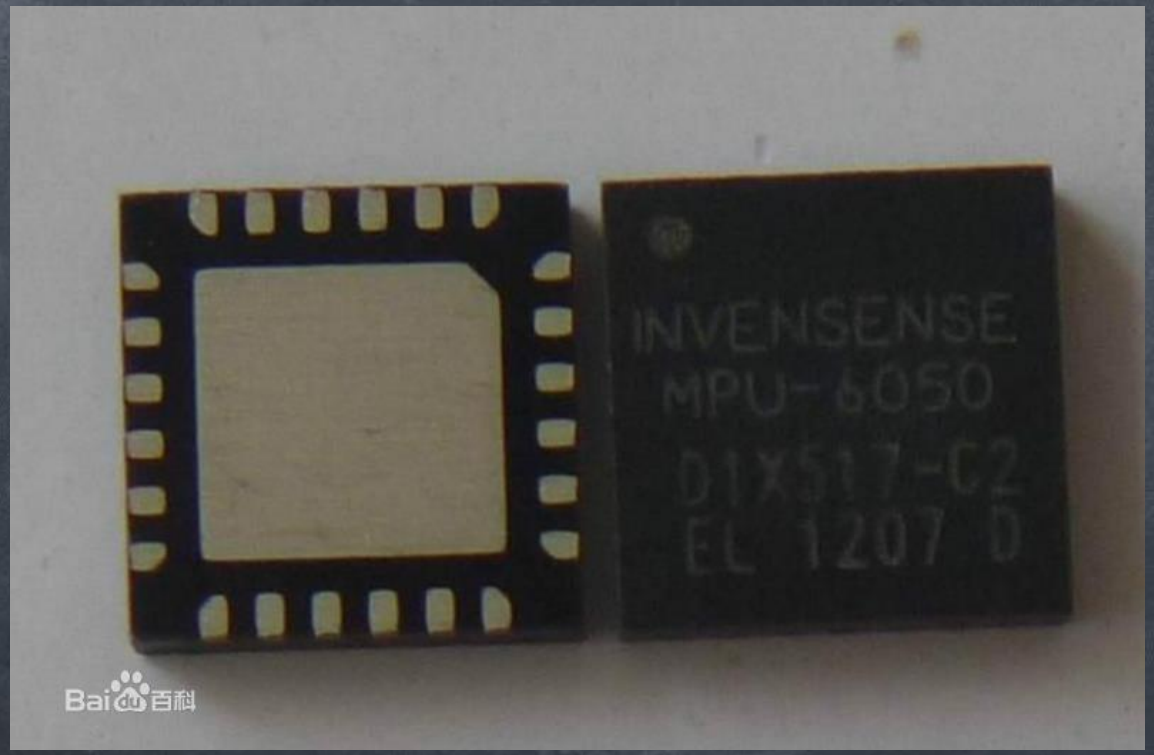
从日常数据中寻找标准数据的算法

用机器学习进行模式匹配

03.课题研究的进展

The Research Progress

技术节点.零
传感器参数及其数据



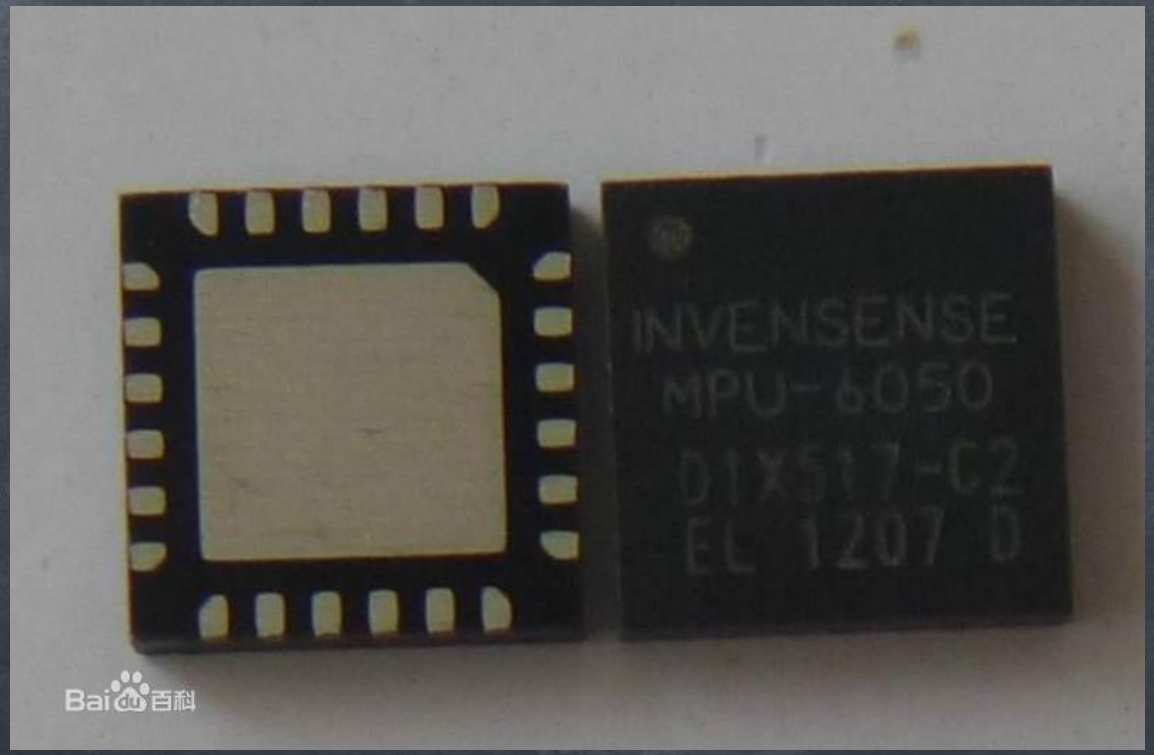
我们在攻击过程中，需要先了解手机的传感器。目前手机所装备的传感器有很多种，精度都不错，咱们以mpu6050举个例，这款传感器相对来说比较低端，但是性能很好。
MPU6050是InvenSense公司推出的一款低成本的6轴传感器芯片，包括三轴加速度，三轴角速度。其体积小巧，用途非常广。做平衡小车，四轴飞行器，飞行鼠标等等。

精度：
16384LSB/g(±2g)

03.课题研究的进展

The Research Progress

技术节点.零
传感器参数及其数据



LSB当测量的加速值是1g（重力加速度）时，那么加速度的输出就是16384

那么为什么是16384呢，我们计算一下：
16384*2=32768，
32768*2=65536=2^16，MPU6050的ADC是16位的，所以不管测量范围多大，最终的输出范围都不会超过65535，所以测量范围越大，精度就越低。下面计算一下测量范围是±16g时的精度：
16*2/65536=0.00048828125，然后取倒数1/0.00048828125=2048，
从这里我们可以看出，16个g是日常绝对达不到的加速度值，但是此时它的精度还可以达到10^-3，所以其精度是很好的。

03.课题研究的进展

The Research Progress

	位置0(0度)			位置1(26.6度)			位置2(45度)			位置3(90度)			测量时间
	tx	ty	tz	tx	ty	tz	tx	ty	tz	tx	ty	tz	
循环1次	-1.4	-2.89	0	-1.33	-3.03	28.88	-1.3	-3.13	46.8	-1.4	-3.31	92.11	
循环2次	-1.49	-2.85	3.43	-1.45	-2.95	30.59	-1.34	-3.1	48.05				
循环3次				-1.45	-2.93	31.79	-1.47	-2.98	50.39	-1.42	-3.48	97.67	
循环4次	-1.44	-2.92	10.65	-1.35	-3.01	36.95	-1.37	-3.26	53.07				710秒
循环1次	-1.5	-3.02	0	-1.47	-3.11	25.29	-1.4	-3.21	43.57	-1.42	-3.42	87.51	
循环2次	-1.54	-2.99	0.52	-1.47	-3.04	24.78	-1.39	-3.17	43.53				
循环3次				-1.46	-3.05	26.07	-1.4	-3.16	43.07	-1.44	-3.41	89.29	350秒
循环1次	-1.58	-3.04	0	-1.47	-3.12	30.45	-1.43	-3.19	48.76	-1.47	-3.41	95.07	
循环2次	-1.57	-2.99	6.51	-1.48	-3.08	32.76	-1.42	-3.17	50.15				
循环3次				-1.47	-3.08	32.79	-1.42	-3.19	41.83	-1.46	-3.4	88.28	306秒

关于mpu6050错误使用的例子

在纸上画出四个角度，分别为0度、26.6度、45度、90度，将纸固定在桌子上，每次将陀螺仪模块旋转到对应的位置，经过多次实验，证实所使用的MPU6050模块的角度输出确实存在很大的误差。但是从datasheet上看，MPU6050芯片和其他精度高且价格昂贵的芯片如ADIS16系列差别不是大，所以可以得出，极有可能是姿态融合算法部分的问题。（这一部分为引用）

03.课题研究的进展

The Research Progress

$$X(k|k-1)=A\,X(k-1|k-1)+B\,U(k) \quad \ldots\ldots\ldots\ldots (1)$$

$$P(k|k-1)=A\,P(k-1|k-1)\,A'+Q \quad \ldots\ldots\ldots (2)$$

$$X(k|k)=X(k|k-1)+Kg(k)\,(Z(k)-H\,X(k|k-1)) \quad \ldots\ldots\ldots (3)$$

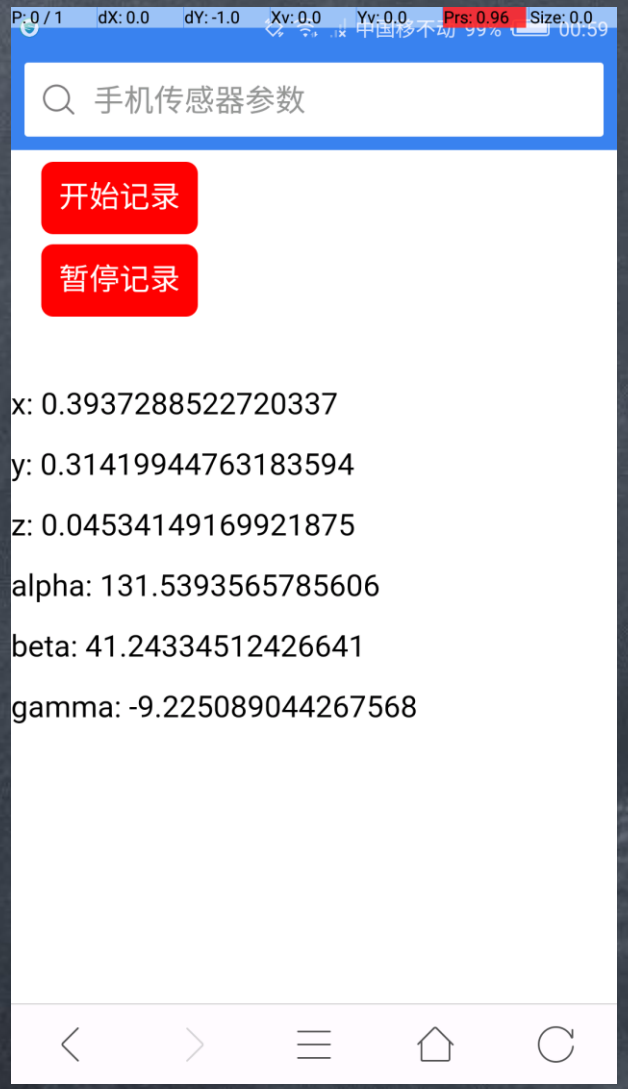
$$Kg(k)=P(k|k-1)\,H'\,/\,(H\,P(k|k-1)\,H'+R) \quad \ldots\ldots\ldots (4)$$

$$P(k|k)=(I-Kg(k)\,H)\,P(k|k-1) \quad \ldots\ldots\ldots (5)$$

解决思路就是，给加速度和角速度不同的权值，把它们结合到一起，进行修正，然后根据上一时刻对这个时刻的预测值和这个时刻的真实值进行对比，多次迭代，调整参数，这就是卡尔曼滤波器

03.课题研究的进展

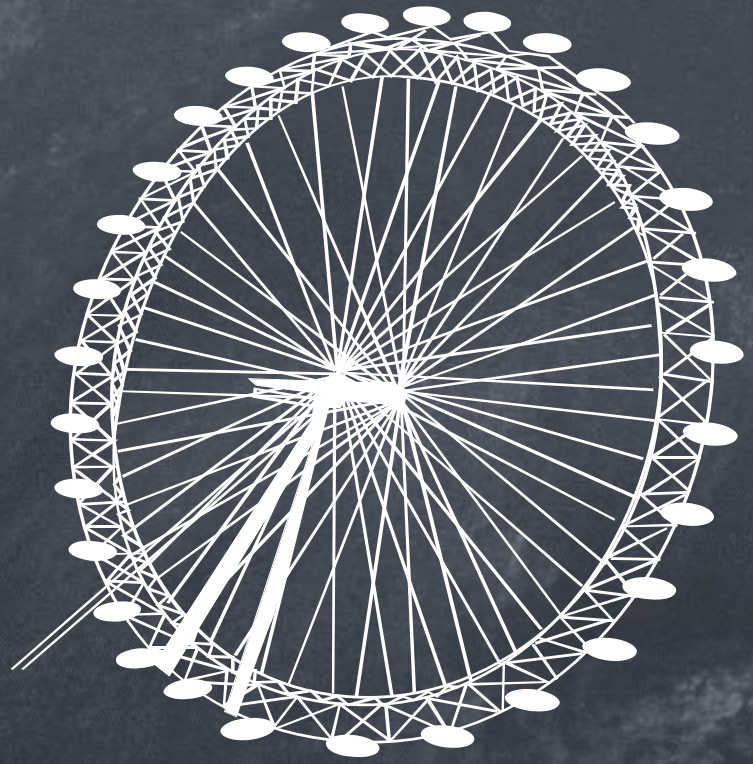
The Research Progress



技术节点.一
获取数据并传输

03.课题研究的进展

The Research Progress

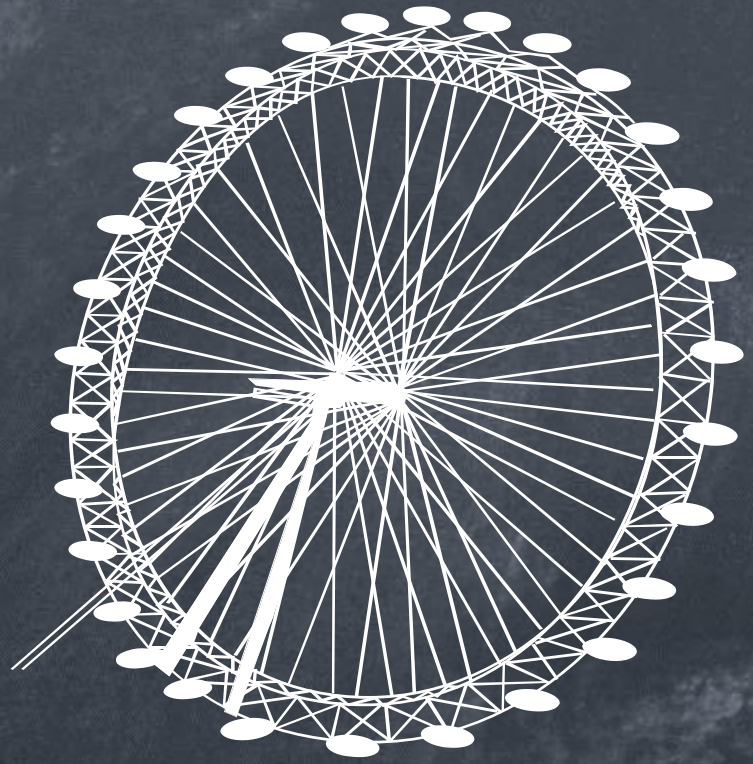


技术节点.二
预处理数据算法

数据采集和预处理的目的都是用来形成训练集以供机器学习，针对不同的特征，数据采集的方式和数据与处理的方式不是一成不变的，如何能够让这些特征在真实的情况下，最大化的反应出来应该是很重要的。所以我在数据采集之前，根据所需要的数据来明确自己需要什么样的数据，采集数据同时就在预处理数据，不断修正。

03.课题研究的进展

The Research Progress



技术节点.二
预处理数据算法

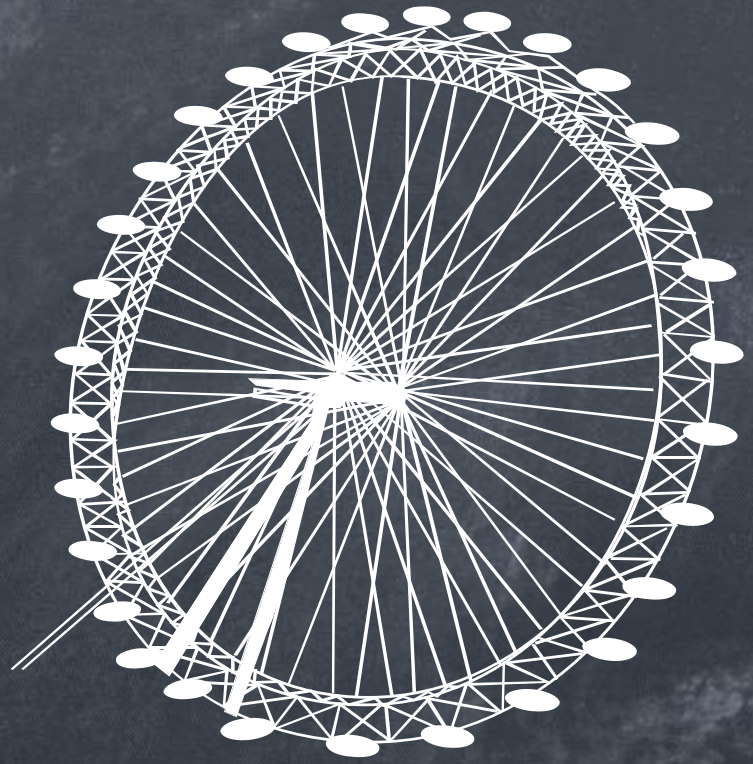
以0-9密码为例，我准备数据库有这些，

1，　　传统的单个按键线性姿态数据。就是在输入单个的0~9十个数字时，手机产生的一系列姿态信息。

2，　　组合按键的线性姿态数据。以6位0~9数字密码来说，将所有排列组合所产生的加速度信号全部采集出来，理论上会有10^6种，且这种数据采集工程量太大，而且不同种类之间的区别很小，所以只可以作为参考。

03.课题研究的进展

The Research Progress



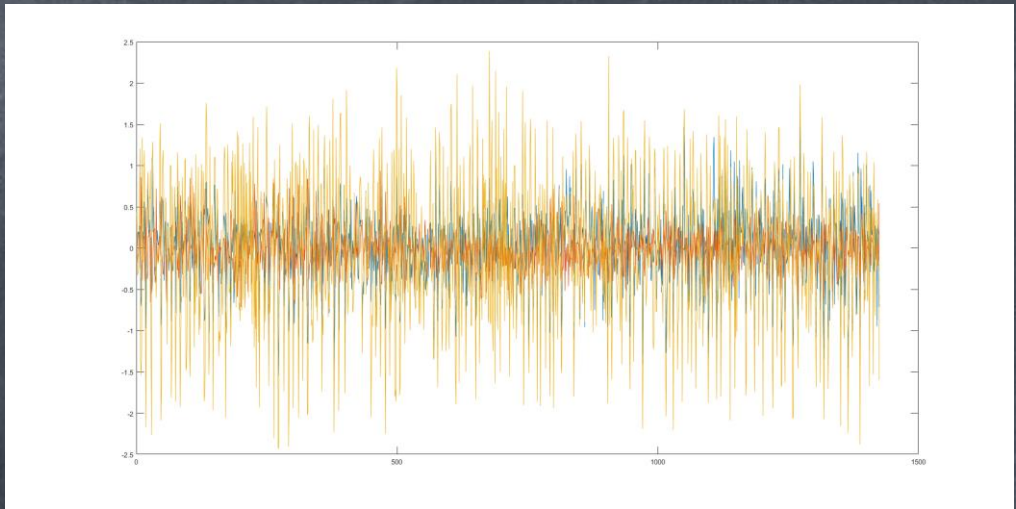
技术节点.二
预处理数据算法

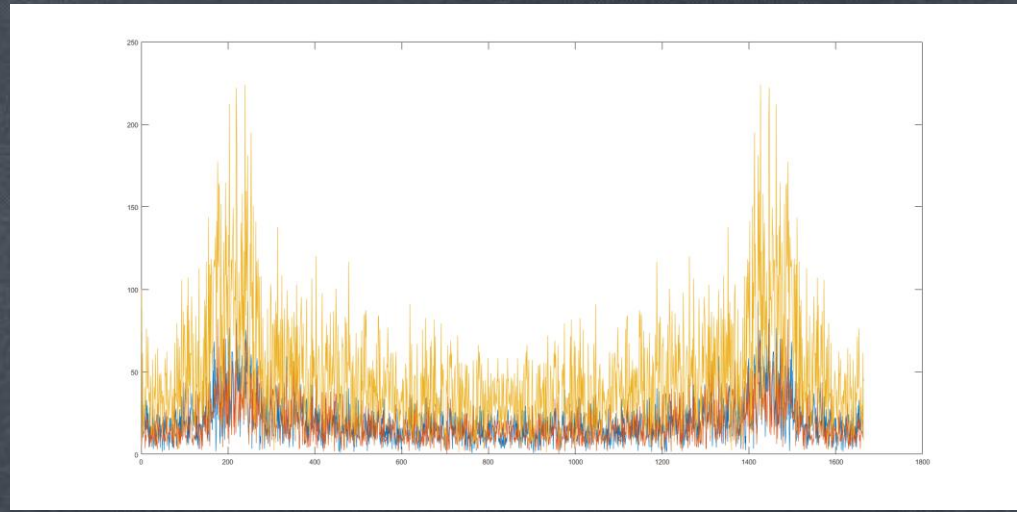
在采集过程中，我发现两个问题，首先，不同用户或者同一用户多次输入时，每次输入时间是不固定的，所以上述涉及到线性信号的采样方式都遇到了挑战。
其次，在训练数据采集时，输入前后都是有其他动作的，这些动作所产生的加速度信号会很大程度上干扰有效的加速度信号，因为输入动作起始点是没有一个明确信号的。

03.课题研究的进展

The Research Progress

技术节点.二
预处理数据算法



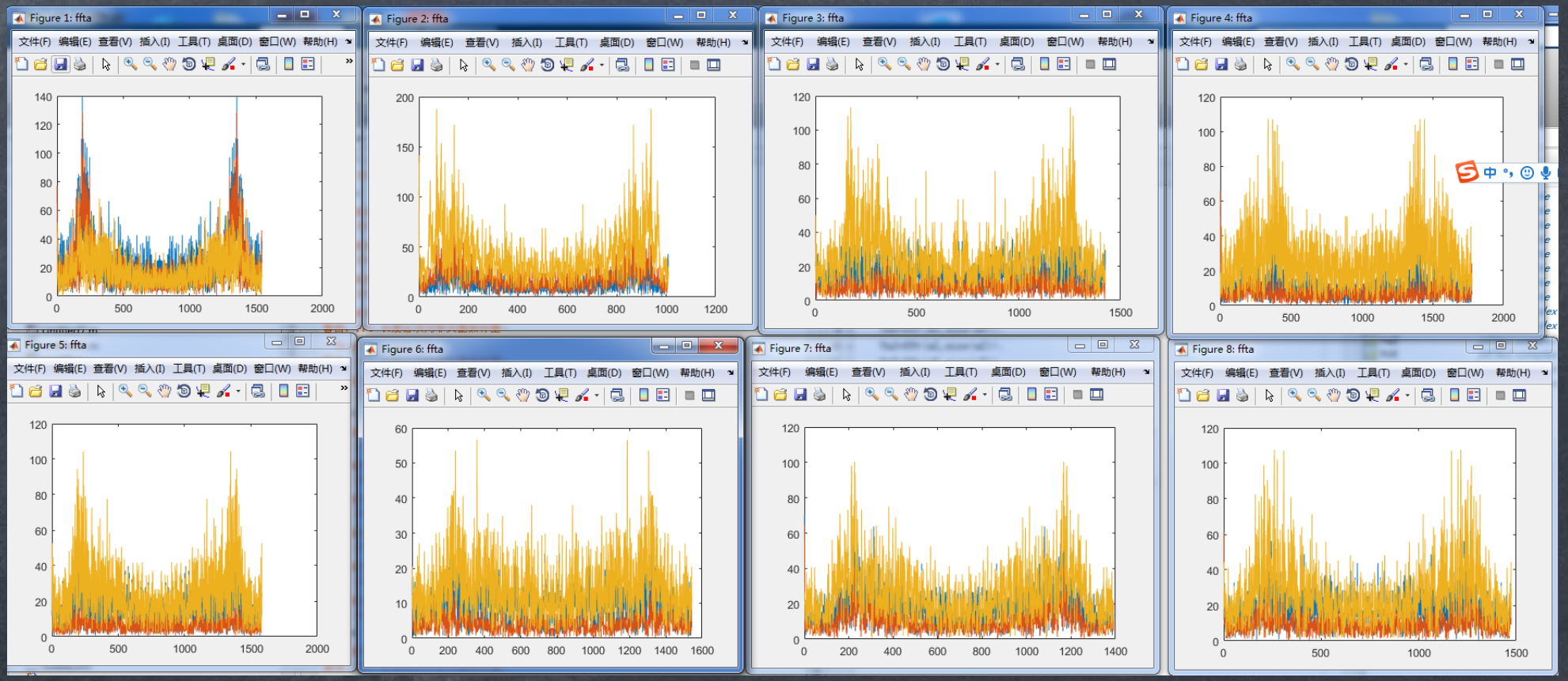


对时域信号进行频域变换变换，并多次重复输入动作， 这样可以将每种动作不同次输入时相同的部分在频域能量分布上更显著的表现出来。同时还大大减少了单词随机不正常输入所带来的影响。

03.课题研究的进展

The Research Progress

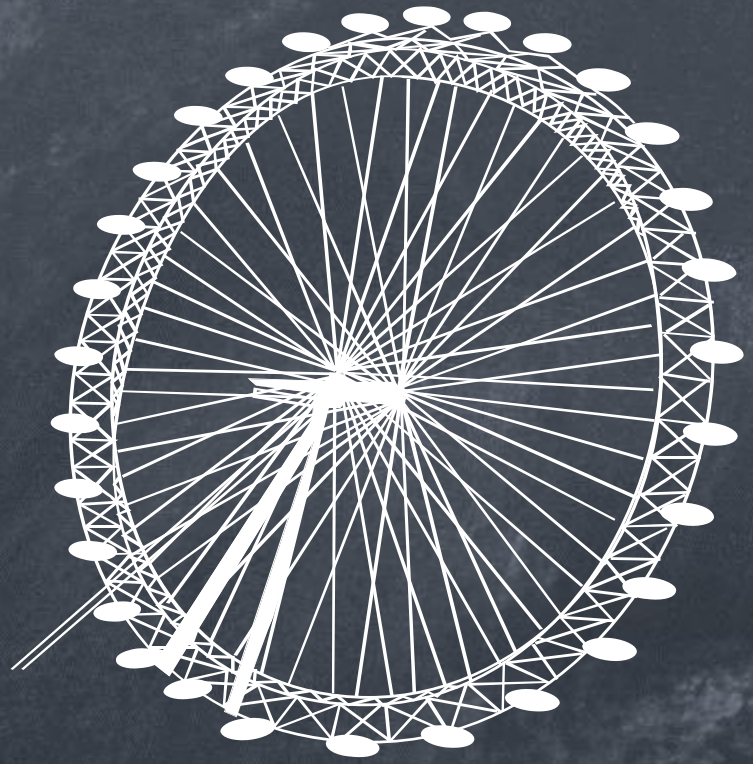
技术节点.三
构建不同模式下的数据库



不同输入文本之间的差异甚至肉眼可辨

03.课题研究的进展

The Research Progress



下一个技术节点就是模式匹配了，有两个方向，可以采用统计模式识别，
这里需要建立很多模型，比如kNN，随机森林，决策树，C4.5算法，支持向量机，贝叶斯分类等等，然后把训练数据全部扔进去，不断调试，根据每个模型的表现打分，选取相对靠谱的选出来，再用一次卡尔曼滤波。
另一个方向比较大胆，就是建立卷积神经网络，利用卷积网络对于图像识别的优势，把生成的波形图直接扔进去，对波形图进行学习分类。

03.课题研究的进展

The Research Progress

然而，
在建立机器学习模型的时候我突然意识到了一个问题。。。。

03.课题研究的进展

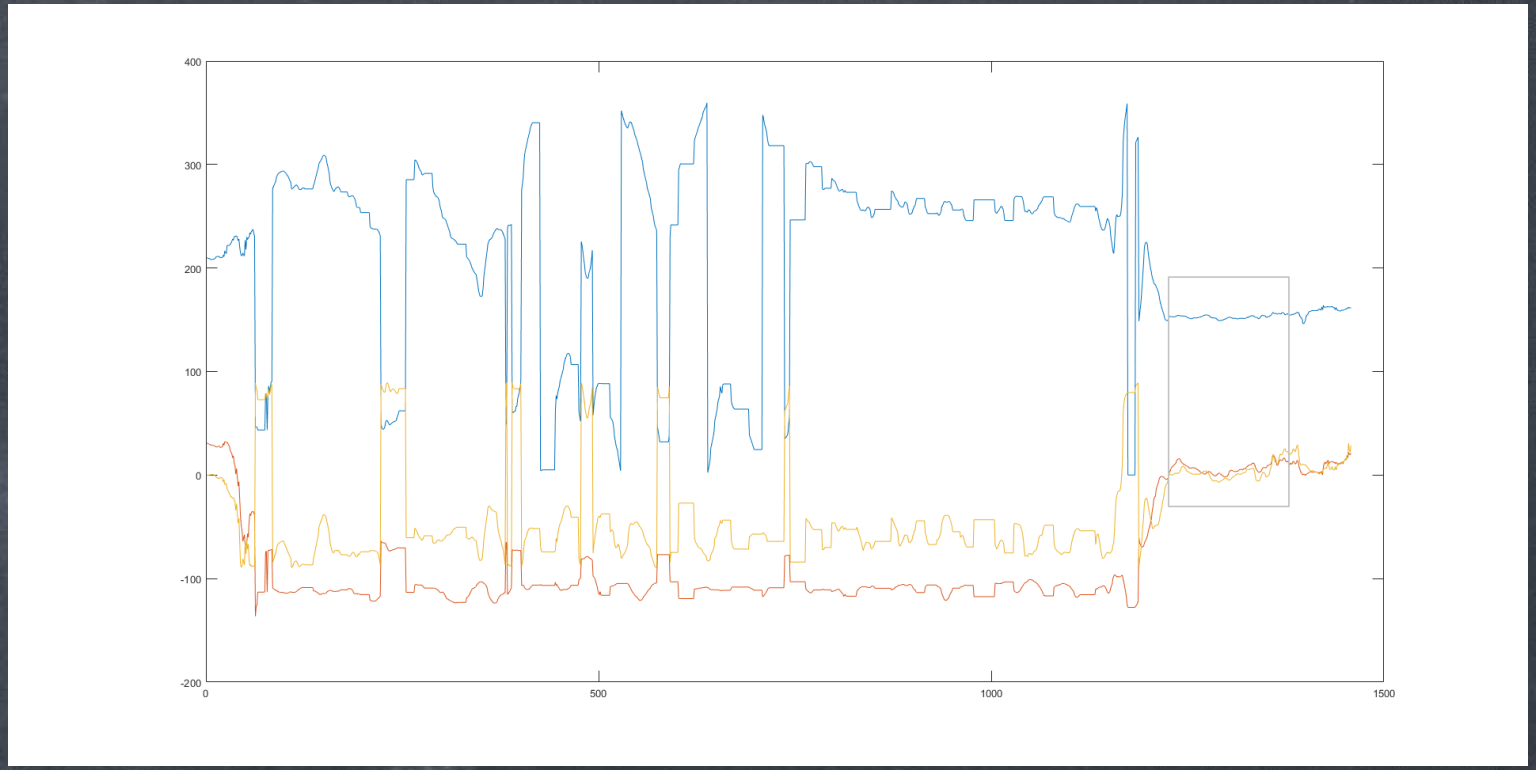
The Research Progress

无论我用来统计，处理，机器学习的，都是制作好的，很规范的数据。然而，实际生活中我们的加速度传感器产生的数据是很杂乱的。如何从大把大把的数据中找到你想要的数据呢？当你拥有获取加速度流的工具，并且成功回传数据之后，如何从这一大段加速度流中找到有价值的数据呢？

03.课题研究的进展

The Research Progress

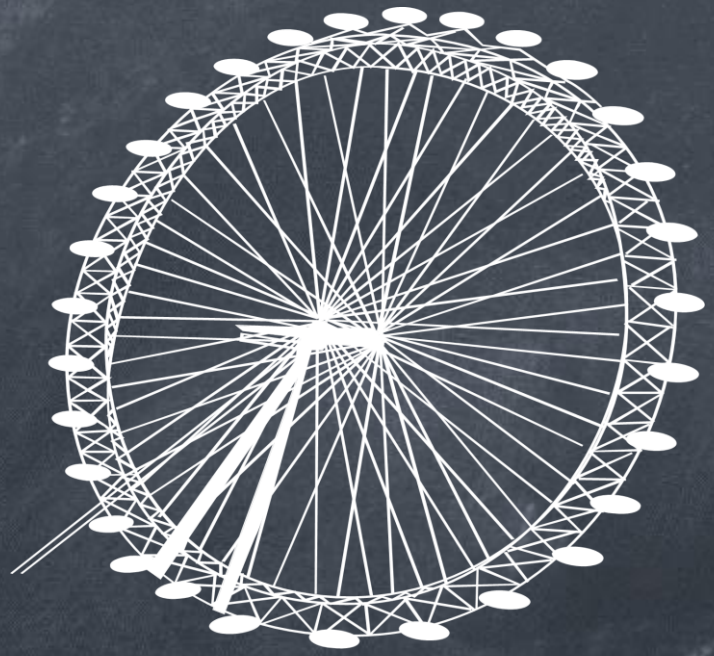
技术节点.四
从日常数据中寻找标准数据的算法



由于网页可以在后持续采集数据所带来的两个问题，1，多余的电量消耗可能会引起用户警觉，2，如何从日常环境中复杂的传感器数据中找到有效的数据进行下一步运算？

03.课题研究的进展

The Research Progress



如果直接对所有数据进行处理，也会导致两个问题，

1，计算效率特别底下，手机息屏的时间相对于用户操作手机的时间，尤其是相对于输入敏感文本的时间来说太长，长的时间就会带来多的无效数据，
2，会干扰数据的处理，在长时间的息屏状态下，很难保证不会产生一些动作的加速度流与输入某些文本的加速度流相似，如果将他们也定义为文本，那么精度会变得很低。

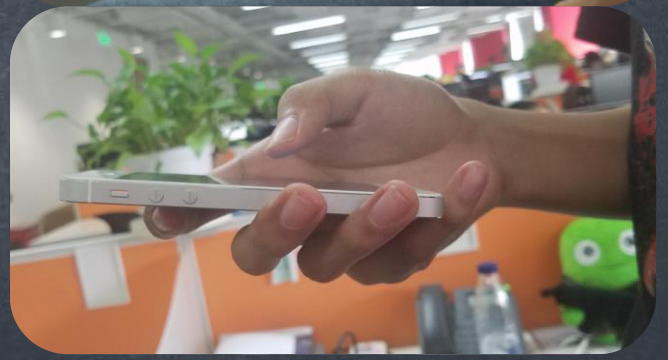
03.课题研究的进展

The Research Progress

技术节点.四
从日常数据中寻找标准数据的算法

由于人体结构，不同姿势使用手机时手机的角度是不同的，尤其在输入文本时，人们更倾向于将手机放到几个相对舒服的姿势。

坐姿使用



站姿使用

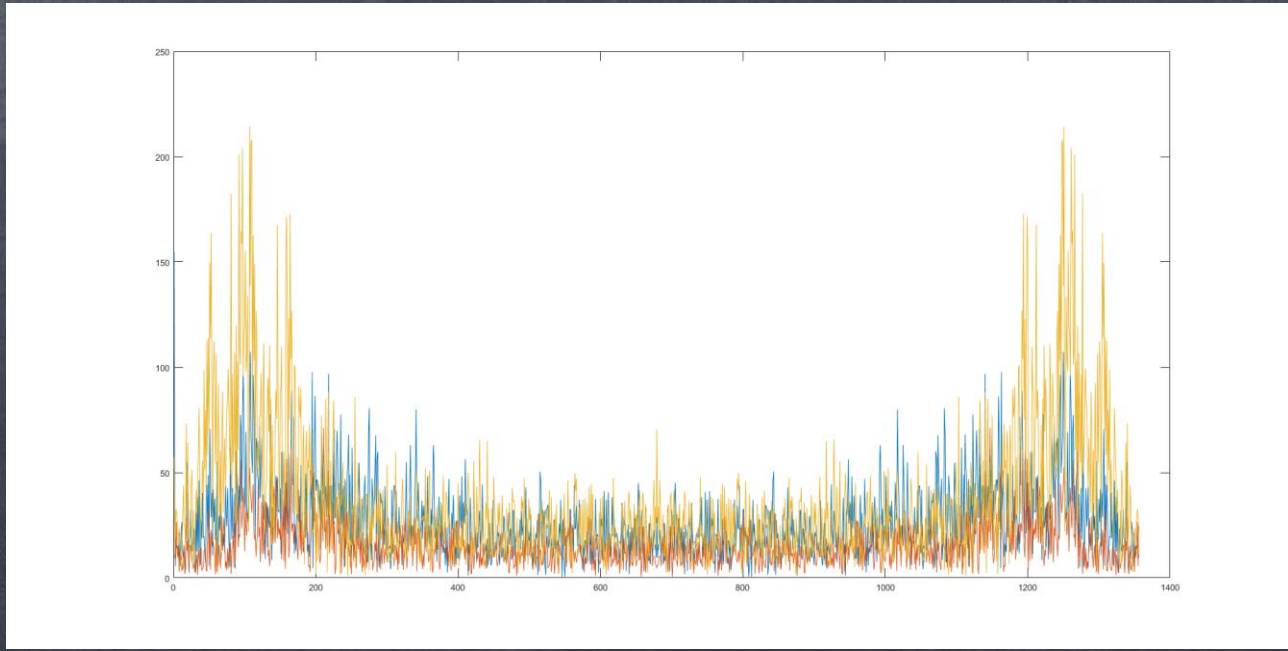
根据不同的使用场景分别建立模型，目前从角度数据可以区分的有五种通过调用电磁传感器可以区分左右手并再次分类

卧姿使用

03.课题研究的进展

The Research Progress

技术节点.四
从日常数据中寻找标准数据的算法



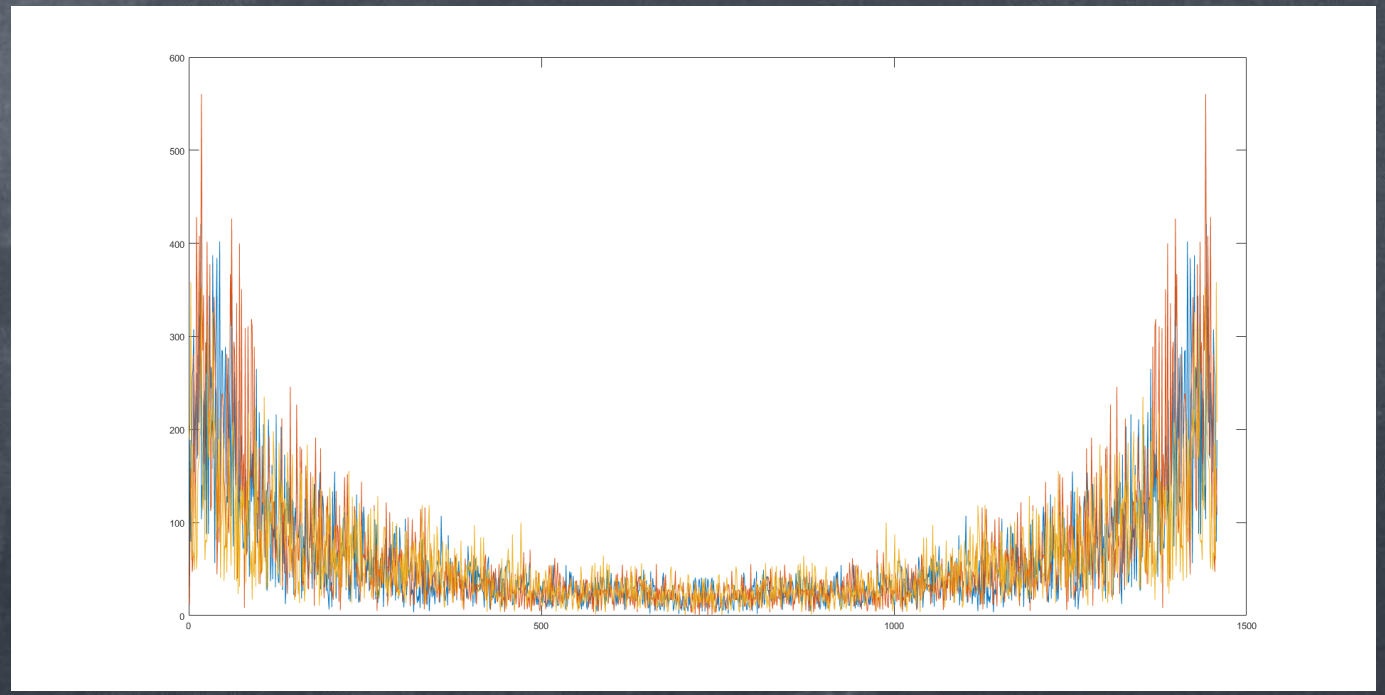
由于我采用的是频域能量分布处理，这样就会产生一个在一定频率内的能量分布，我们只能通过大量的运算来推测这些能量可以由输入几个“1”，几个“2”来组成。最后得出的结果可能是
[2，2，8，9，0，0]

这组数据中，模型只能确定有[1,2,3,4,5,6]没有顺序，需要多次排列组合测试

03.课题研究的进展

The Research Progress

技术节点.四
从日常数据中寻找标准数据的算法

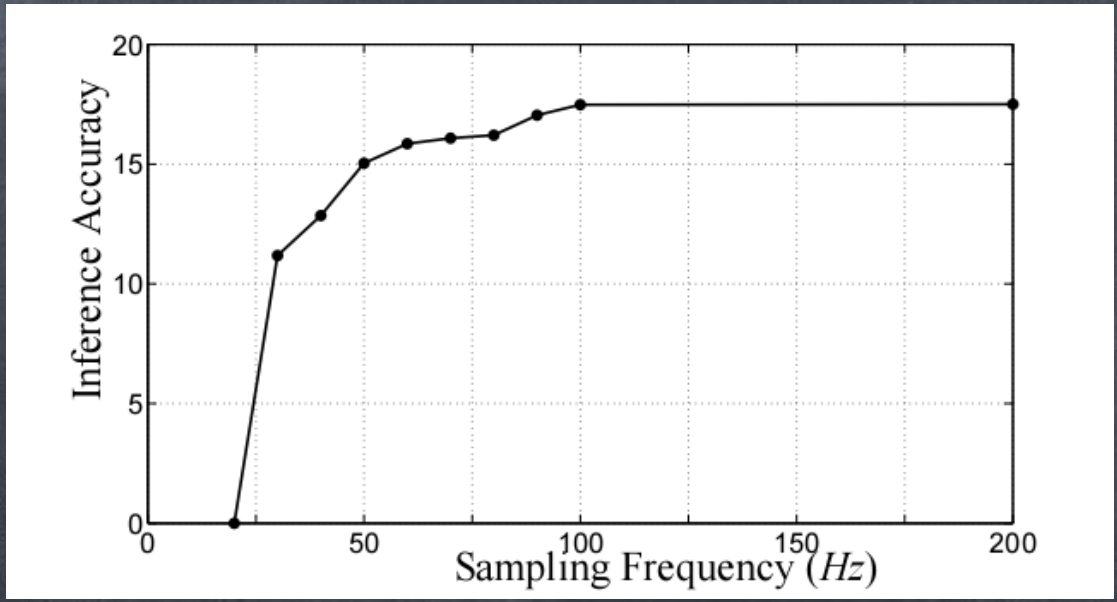


图为一个主对角线滑动所产生的能量信号，若仅有0-9数字作为分类模型，无法拟合出这个能量分布

同时，更坏的或者说是更可能出现的情况是会出现很多组预测数列组。而且，这样做还需要将整个屏幕进行画格区分而不只是针对输入键盘的九宫格因为无法确定哪些是输入操作哪些是无关操作，所以只能对全屏进行统计，来找全屏幕的输入分布，再根据布局将输入位置映射下来。又由于输入背景是无法观察的，所以这种方式仅是理论上可行。举个例子，一个很简单的滑动操作就可以让这个预测系统崩溃。

03.课题研究的进展

The Research Progress



在试图检索资料解决这个问题时，我发现了已经有论文指出：

人为的降低传感器采样频率可以有效地减少可能的，由传感器信息泄露而导致的风险

03.课题研究的进展

The Research Progress

然而我们独角兽的实习生绝不认输！！

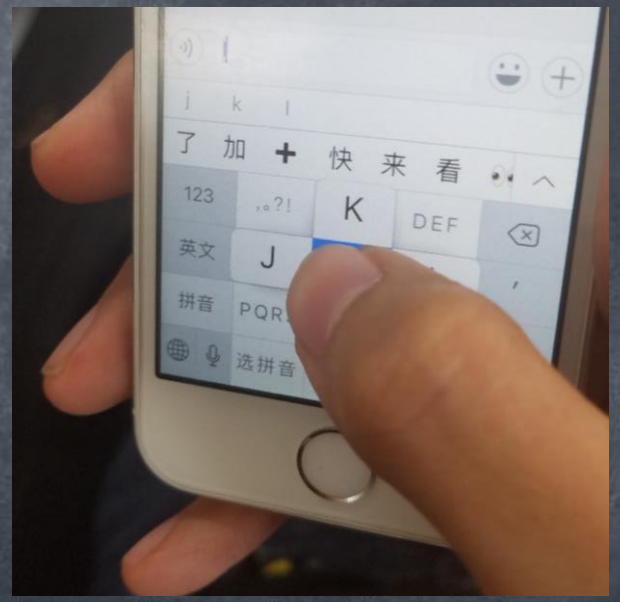
03.课题研究的进展

The Research Progress

机器学习下的关键在数据，限制采样率就是无法获得足够的数据，实战是在缺乏数据的情况下进行的，那么在选择训练数据时，就不能用数据很充分的情况进行训练，否则不具备实际意义。具体来说就是利用50hz采样率数据训练出来的模型，对于20hz采样率的数据，是很难或者说无法有效的。

03.课题研究的进展

The Research Progress



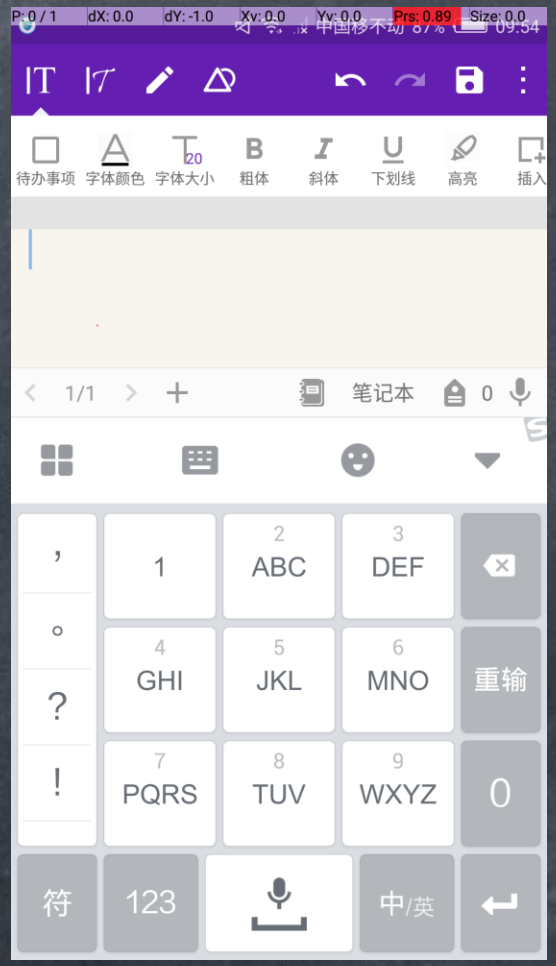
一个左移动作

以多个相对简单的模式识别，来代替一个复杂的模式识别。打个比方来说，就像一个文学素养不是太高的人，无法找到一个精确的成语来描述一个事物，但是它可以通过多个相对简单的形容词来描述，一样可以达到交流的目的。

新建两个数据库，一个是连续的输入动作识别库，如图
另一个是离散的输入姿态数据库，这个数据库只记录有效触屏瞬间手机的姿态信息

03.课题研究的进展

The Research Progress



什么叫做有效触屏？

以搜狗键盘为例，单个按键为圆角矩形，宽为224，高为212，垂直间隔为17，水平间隔为15

按键1，纵向1020pixel~1232pixel，横向187pixel~412pixel
按键2，纵向1020pixel~1232pixel，横向427pixel~652pixel
按键3，纵向1020pixel~1232pixel，横向667pixel~892pixel
按键4，纵向1249pixel~1461pixel，横向187pixel~412pixel
按键5，纵向1249pixel~1461pixel，横向427pixel~652pixel
按键6，纵向1249pixel~1461pixel，横向667pixel~892pixel
按键7，纵向1478pixel~1690pixel，横向187pixel~412pixel
按键8，纵向1478pixel~1690pixel，横向427pixel~652pixel
按键9，纵向1478pixel~1690pixel，横向667pixel~892pixel
按键0，纵向1707pixel~1903pixel，横向427pixel~652pixel

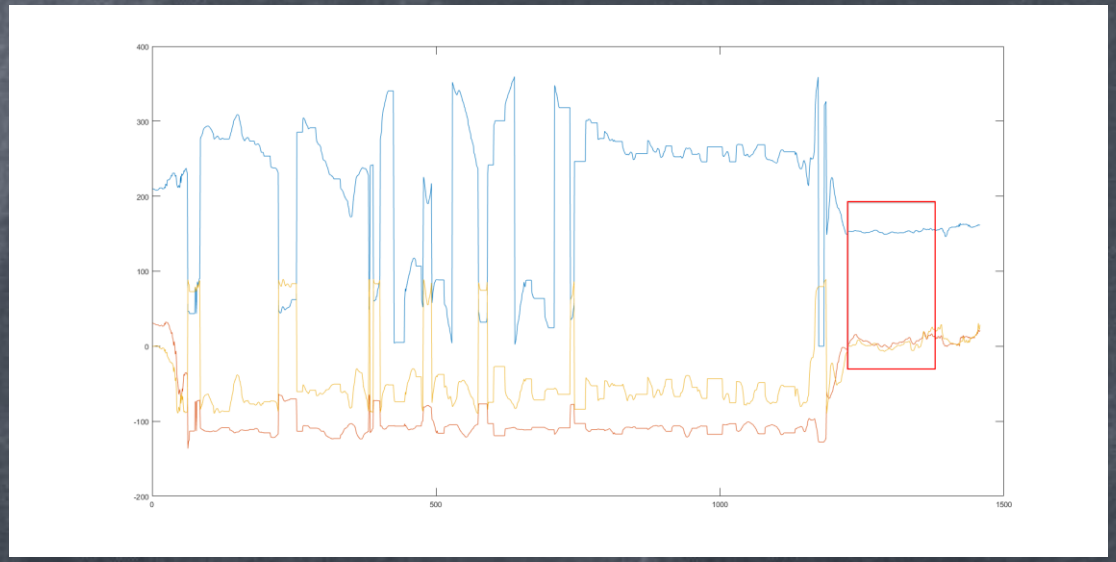
以上为有效区域。落在有效区域内的触摸动作为有效触屏

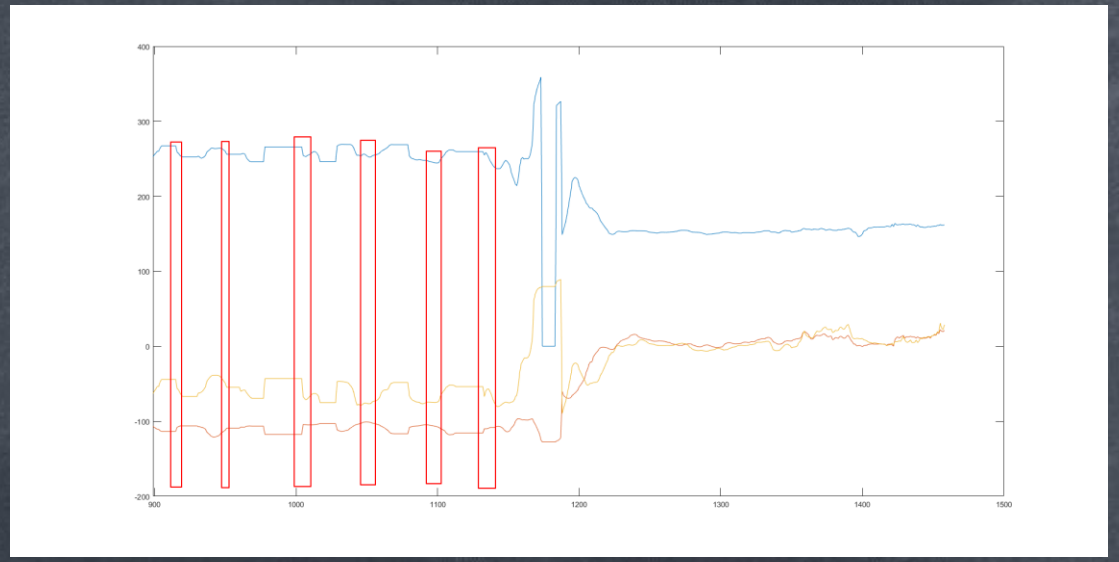
我已经研究了市面上几个主流的输入键盘作为数据库。

03.课题研究的进展

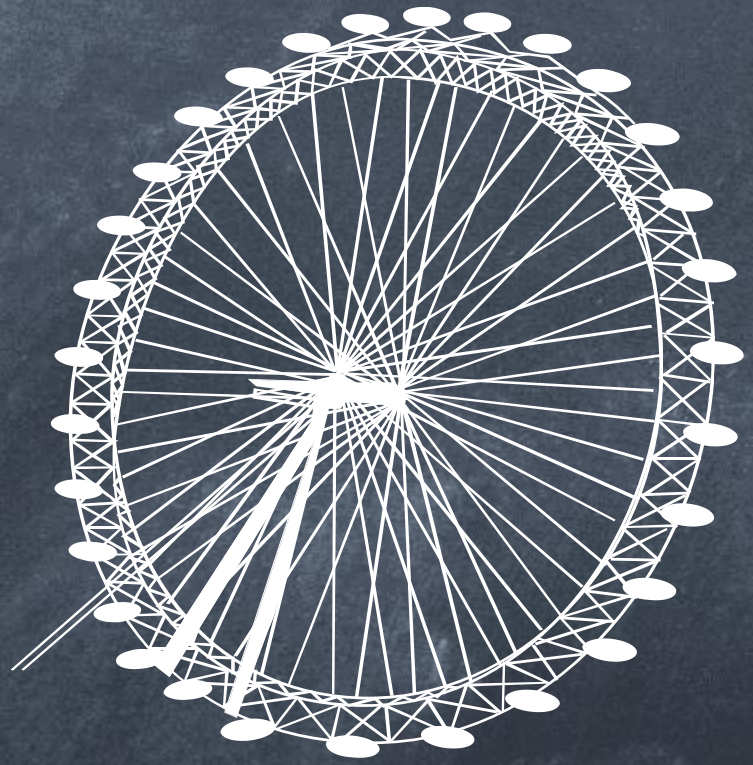
The Research Progress

技术节点.四
从日常数据中寻找标准数据的算法





先线性处理，再频域处理。先在时域寻找这些特定的姿态，如果某个时间段内有连续的特定姿态的变换，那么可以确认这一段时间在进行文本输入，将其提取出来，在其附近小范围内进行频域能量分析，那么就可以有效地从复杂的实际使用中找到我们需要的“输入文本时”的传感器数据，然后在进行下一步运算。同时，在线性处理过程中，就会产生一个对于输入文本的预测，这个预测仅来自于手机姿态，但精度不高，凭借其分布状况，可以确认某一时段“是否在输入字”，并将那个文本预测作为参考。这样分类就可以达到一个可以辨认的层次。



Part 04

课题实践的可行性与展望

The Difficulty Of Practice

04.课题实践的可行性与展望

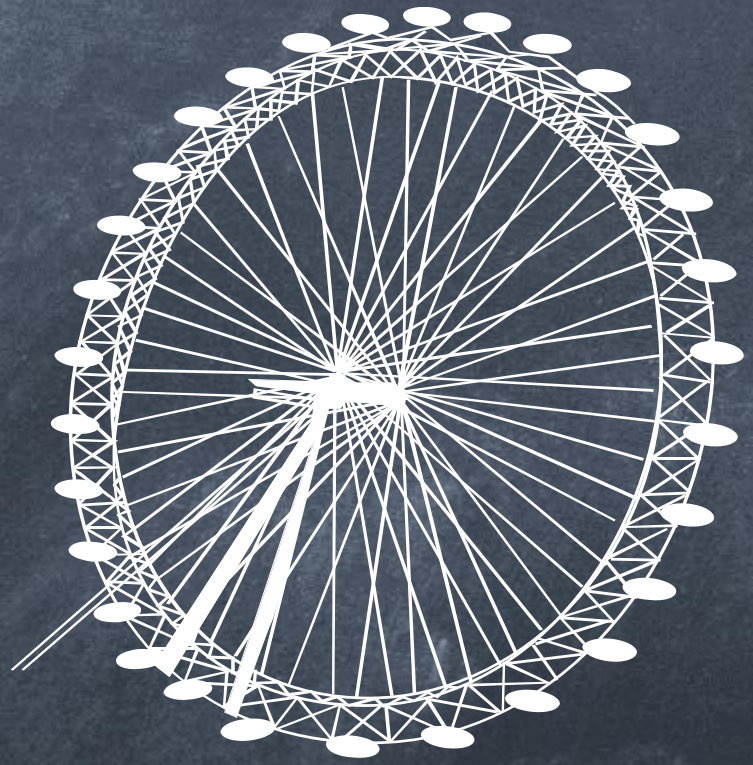
The Difficulty Of Practice

数据量是否足够大

机器学习模型是否可以进一步优化

攻击网页如何进行伪装

能否根据语言学从输入文本中找到敏感信息



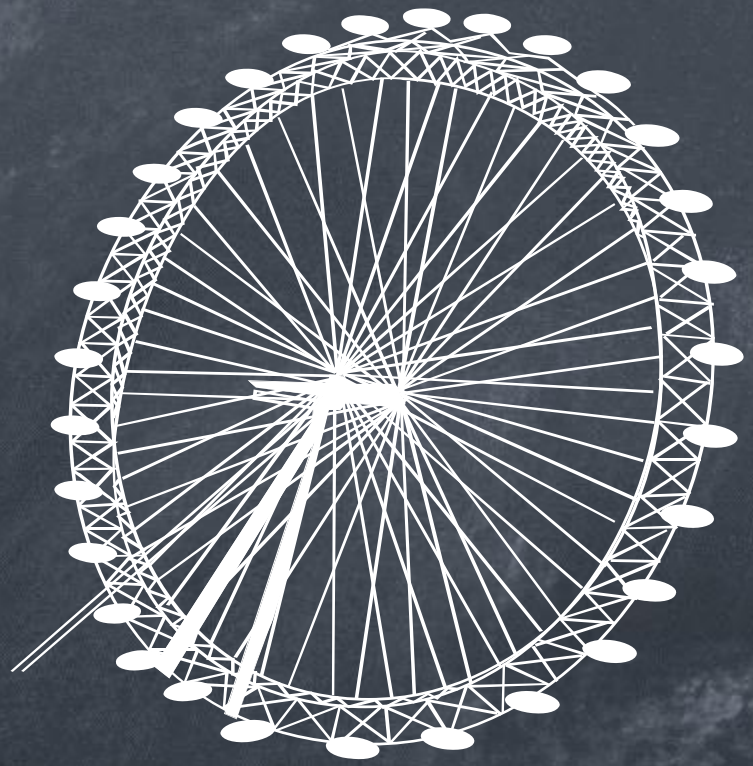
Part 05

课题研究的结论

The Conclusion Of Study

05.课题研究的结论

The Conclusion Of Study



通过传感器来预测输入文本是旁信道攻击的一种，在目前并不十分热门，但是我认为这种攻击方式可以在不利用传统漏洞的同时，大大增大预测的精确度，可以与其他攻击手段一同进行，而且难以防治，虽然目前我的研究还不够成熟，但是对于这个想法，我将一直持续研究下去。而我制作的两个工具我也将放到网上，供对此方向有兴趣的人研究使用，网页近期也将会放到阿里云上，面向更多的人来采集更多的数据。

THANK YOU

欢迎各位老师指导